CLAIMS

WHAT IS CLAIMED IS:

1. A method for predicting a polyadenylation site comprising:

inputting a plurality of RNA transcript sequences or sequences dervied from RNA transcript sequences, wherein at least one sequence has its poly A or poly T tract sequence;

searching for a polyadenylation site, wherein the polyadenylation is an adenine rich region at the end of the sequence or a thymine rich region at the beginning of the sequence;

detecting the presence of polyadenylation signals neigboring the polyadenylation site by scanning the EST or RNA sequences or their corresponding genomic DNA sequences.

2. The method of Claim 1 wherein the step of searching for a polyadenylation site comprising scanning the sequences for adenine rich region at the end of the sequence or a thymine rich region at the begining of the sequence.

3. The method of Claim 2 wherein the adenine rich region comprises adenine in at least 50% of the region and the thymine rich region comprises thymine in at least 50% of the region.

4. The method of Claim 2 wherein the adenine rich region comprises adenine in at least 60% of the region and the thymine rich region comprises thymine in at least 60% of the region.

5. The method of Claim 2 wherein the adenine rich region comprises adenine in at least 70% of the region and the thymine rich region comprises thymine in at least 70% of the region.

6. The method of Claim 2 wherein the adenine rich region comprises adenine in at least 80% of the region and the thymine rich region comprises thymine in at least 80% of the region.

7. The method of Claim 1 wherein a heuristic score $n_A / (n_A + 0.5*(\max(n_R-20,0)))$ is used for detecting adenine or thymine rich region; wherein n_A is the number of adenines or thymines in the block, and n_R is the number of bases after the block of adenines or thymine to the end of the sequence.

8. A method for detecting polyadenylation signal in a sequence with a polyadenylation site comprising searching for a polyadenylation signal hexamer in the sequence before the polyadenylation.

9. The method of Claim 8 wherein the searching comprises evaluating the probability that there is a polyadenylation site: $Pr(h=k|x)$ for $k = 6,7,...,N$, wherein the sequence before the polyadenylation site is $x=(x_1,x_2,...x_N)$ and where x_N is the 3'-most base before the polyadenylation site.

10. The method of Claim 9 wherein: $Pr(h=k|x) = Pr(x|h=k)\ Pr(h=k)/Pr(x)$.

11. The method of Claim 10 wherein $Pr(h=k|x) = Pr(x_{k-5},...,x_k|h=k)\ Pr(h=k)/Pr(x_k-5,...,x_k)$ and wherein $Pr(h=k)$ is the probability that the polyadenylation hexamer is located at position k in the sequence, at a distance (N-k) from the polyadenylation site, $Pr(x_{k-5},...,x_k|h=k)$ is the probability of observing the hexamer $(x_{k-5},...,x_k)$ given that it is a polyadenylation signal and $Pr(x_{k-5},...,x_k|h \neq k)$ is the probability of observing the hexamer given that it is not from a polyadenylation signal.

12. The method of Claim 11 wherein the step of detecting comprises using a gamma function to produce a density which places the majority of its weight on the positions located 5 to 25 bases distant from the polyadenylation site.

13. The method of Claim 12 wherein $Pr(x_{k-5},...,x_k|h \neq *)$, the probability of observing the hexamer given that it is not from a polyadenylation signal, is modeled using a second-order Markov model trained on data collected from human 3' UTRs.

14. The method of Claim 13 wherein $Pr(x_{k-5},...,x_k|h \neq *)=Pr(x_{k-5})\ Pr(x_{k-4}|x_{k-5})\ Pr(x_{k-3}|x_{k-5},x_{k-4})\ Pr(x_{k-2}|x_{k-4},x_{k-3})\ Pr(x_{k-1}|x_{k-3},x_{k-2})\ Pr(x_k|x_{k-2},x_{k-1})$, wherein the first term is a zero-order Markovian probability, the second is a first-order Markovian probability and the remaining four terms are second-order Markovian probabilities.

15. The method of Claim 14 wherein, for a k^{th}-order Markov model, the probability of base b following a word w of length k is estimated by the frequency of the concatenated word (wb) divided by the frequency of the word w, where frequencies are computed from the training dataset of 3'UTR sequences.

16. The method of Claim 15 wherein, for the case k=0 (a zero-order Markovian model), the probability of base b is estimated by its frequency in the dataset divided by the size of the dataset.

17. A computer readable medium comprising computer-executable instructions for performing the method comprising:

inputting a plurality of RNA transcript sequences or sequences dervied from RNA transcript sequences, wherein at least one sequence has its poly A or poly T tract sequence;

searching for a polyadenylation site, wherein the polyadenylation is an adenine rich region at the end of the sequence or a thymine rich region at the beginning of the sequence;

detecting the presence of polyadenylation signals neigboring the polyadenylation site by scanning the EST or RNA sequences or their corresponding genomic DNA sequences.

18. The computer readable medium of Claim 17 wherein the step of searching for a polyadenylation site comprising scanning the sequences for adenine rich region at the end of the sequence or a thymine rich region at the begining of the sequence.

19. The computer readable medium of Claim 18 wherein the adenine rich region comprises adenine in at least 50% of the region and the thymine rich region comprises thymine in at least 50% of the region.

20. The computer readable medium of Claim 19 wherein the adenine rich region comprises adenine in at least 60% of the region and the thymine rich region comprises thymine in at least 60% of the region.

21. The computer readable medium of Claim 20 wherein the adenine rich region comprises adenine in at least 70% of the region and the thymine rich region comprises thymine in at least 70% of the region.

22. The computer readable medium of Claim 21 wherein the adenine rich region comprises adenine in at least 80% of the region and the thymine rich region comprises thymine in at least 80% of the region.

23. The computer readable medium of Claim 17 wherein a heuristic score $n_A / (n_A + 0.5*(\max(n_R-20,0)))$ is used for detecting adenine or thymine rich region; wherein n_A is the number of adenines or thymines in the block, and n_R is the number of bases after the block of adenines or thymine to the end of the sequence.

24. A computer readable medium comprising computer-executable instructions for performing the method comprising: searching for a polyadenylation signal hexamer in the sequence before the polyadenylation.

25. The computer readable medium of Claim 24 wherein the searching comprises evaluating the probability that there is a polyadenylation site: $Pr(h=k|x)$ for k = 6,7,...,N, wherein the sequence before the polyadenylation site is $x=(x_1,x_2,...x_N)$ and where x_N is the 3'-most base before the polyadenylation site.

26. The computer readable medium of Claim 25 wherein: $Pr(h=k|x) = Pr(x|h=k)$ $Pr(h=k)/Pr(x)$.

27. The computer readable medium of Claim 26 wherein: $Pr(h=k|x) = Pr(x_{k-5},...,x_k|h=k)$ $Pr(h=k)/Pr(x_{k-5},...,x_k)$ and wherein $Pr(h=k)$ is the probability that the polyadenylation hexamer is located at position k in the sequence, at a distance (N-k) from the polyadenylation site, $Pr(x_{k-5},...,x_k|h=k)$ is the probability of observing the hexamer $(x_{k-5},...,x_k)$ given that it is a polyadenylation signal and $Pr(x_{k-5},...,x_k|h \neq k)$ is the probability of observing the hexamer given that it is not from a polyadenylation signal.

28. The computer readable medium of Claim 27 wherein the step of detecting comprises using a gamma function to produce a density which places the majority of its weight on the positions located 5 to 25 bases distant from the polyadenylation site.

29. The computer readable medium of Claim 28 wherein $Pr(x_{k-5},...,x_k|h \neq k)$, the probability of observing the hexamer given that it is not from a polyadenylation signal, is modeled using a second-order Markov model trained on data collected from human 3' UTRs.

30. The computer readable medium of Claim 29 wherein $Pr(x_{k-5},...,x_k|h \neq k)=Pr(x_{k-5})$ $Pr(x_{k-4}|x_{k-5})$ $Pr(x_{k-3}|x_{k-5},x_{k-4})$ $Pr(x_{k-2}|x_{k-4},x_{k-3})$ $Pr(x_{k-1}|x_{k-3},x_{k-2})$ $Pr(x_k|x_{k-2},x_{k-1})$, wherein the first term is a zero-order Markovian probability, the second is a first-order Markovian probability and the remaining four terms are second-order Markovian probabilities.

31. The computer readable medium of Claim 30 wherein, for a k^{th}-order Markov model, the probability of base b following a word w of length k is estimated by the frequency of the concatenated word (wb) divided by the frequency of the word w, where frequencies are computed from the training dataset of 3'UTR sequences.

32. The computer readable medium of Claim 31 wherein, for the case k=0 (a zero-order Markovian model), the probability of base b is estimated by its frequency in the dataset divided by the size of the dataset.

33. A system comprising: a processor; and a memory coupled with the processor, the memory storing a plurality of machine instructions that cause the processor to perform logical steps of the method comprising:

inputting a plurality of RNA transcript sequences or sequences dervied from RNA transcript sequences, wherein at least one sequence has its poly A or poly T tract sequence;

searching for a polyadenylation site, wherein the polyadenylation is an adenine rich region at the end of the sequence or a thymine rich region at the beginning of the sequence;

detecting the presence of polyadenylation signals neigboring the polyadenylation site by scanning the EST or RNA sequences or their corresponding genomic DNA sequences.

34. The system of Claim 33 wherein the step of searching for a polyadenylation site comprising scanning the sequences for adenine rich region at the end of the sequence or a thymine rich region at the begining of the sequence.

35. The system of Claim 34 wherein the adenine rich region comprises adenine in at least 50% of the region and the thymine rich region comprises thymine in at least 50% of the region.

36. The system of Claim 35 wherein the adenine rich region comprises adenine in at least 60% of the region and the thymine rich region comprises thymine in at least 60% of the region.

37. The system of Claim 36 wherein the adenine rich region comprises adenine in at least 70% of the region and the thymine rich region comprises thymine in at least 70% of the region.

38. The system of Claim 37 wherein the adenine rich region comprises adenine in at least 80% of the region and the thymine rich region comprises thymine in at least 80% of the region.

39. The system of Claim 33 wherein a heuristic score $n_A / (n_A + 0.5*(\max(n_R-20,0)))$ is used for detecting adenine or thymine rich region; wherein: n_A is the number of adenines or thymines in the block, and n_R is the number of bases after the block of adenines or thymine to the end of the sequence.

40. A system comprising a processor; and a memory coupled with the processor, the memory storing a plurality of machine instructions that cause the processor to perform logical steps of the method for detecting polyadenylation signal in a sequence with a polyadenylation site comprising: searching for a polyadenylation signal hexamer in the sequence before the polyadenylation.

41. The system of Claim 40 wherein the searching comprises evaluating the probability that there is a polyadenylation site: $Pr(h=k|x)$ for $k = 6,7,...,N$, wherein the sequence before the polyadenylation site is $x=(x_1,x_2,...x_N)$ and where x_N is the 3'-most base before the polyadenylation site.

42. The system of Claim 41 wherein: $Pr(h=k|x) = Pr(x|h=k)\ Pr(h=k)/Pr(x)$.

43. The system of Claim 42 wherein $Pr(h=k|x) = Pr(x_{k-5},...,x_k|h=k)\ Pr(h=k)/Pr(x_{k-5},...,x_k)$ and wherein $Pr(h=k)$ is the probability that the polyadenylation hexamer is located at position k in the sequence, at a distance (N-k) from the polyadenylation site, $Pr(x_{k-5},...,x_k|h=k)$ is the probability of observing the hexamer $(x_{k-5},...,x_k)$ given that it is a polyadenylation signal and $Pr(x_{k-5},...,x_k|h \neq k)$ is the probability of observing the hexamer given that it is not from a polyadenylation signal.

44. The system of Claim 43 wherein the step of detecting comprises using a gamma function to produce a density which places the majority of its weight on the positions located 5 to 25 bases distant from the polyadenylation site.

45. The system of Claim 44 wherein $Pr(x_{k-5},...,x_k|h \neq k)$, the probability of observing the hexamer given that it is not from a polyadenylation signal, is modeled using a second-order Markov model trained on data collected from human 3' UTRs.

46. The system of Claim 45 wherein $Pr(x_{k-5},...,x_k|h \neq k) = Pr(x_{k-5})\ Pr(x_{k-4}|x_{k-5})\ Pr(x_{k-3}|x_{k-5},x_{k-4})\ Pr(x_{k-2}|x_{k-4},x_{k-3})\ Pr(x_{k-1}|x_{k-3},x_{k-2})\ Pr(x_k|x_{k-2},x_{k-1})$, wherein the first term is a zero-order Markovian probability, the second is a first-order Markovian probability and the remaining four terms are second-order Markovian probabilities.

47. The system of Claim 46 wherein, for a k^{th}-order Markov model, the probability of base b following a word w of length k is estimated by the frequency of the concatenated word (wb) divided by the frequency of the word w, where frequencies are computed from the training dataset of 3'UTR sequences.

48. The system of Claim 47 wherein, for the case k=0 (a zero-order Markovian model), the probability of base b is estimated by its frequency in the dataset divided by the size of the dataset.